GHAJAR
EXHIBIT 37

IN THE UNITED STATES DISTRICT COURT

FOR THE NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADRY, et al,	CASE NO.
	3:23-cv-03417-VC
Plaintiff,	
-against-	
META PLATFORMS, INC.	
Defendant.	

HIGHLY CONFIDENTIAL

UNDER THE PROTECTIVE ORDER

VIDEO-RECORDED DEPOSITION OF

MICHAEL CLARK 30(B)(6)

Zoom Video Conferencing

12/19/2024

3:06 p.m. (MST)

REPORTED BY: MONIQUE CABRERA

DIGITAL EVIDENCE GROUP

1730 M. Street, NW, Suite 812

Washington, D.C. 20036

(202) 232-0646

way more effective to put other mitigations in place earlier in the process from a pretraining perspective in how the model was trained and then also to measure from an ongoing perspective memorization and the likelihood that content could be repeated from training data versus trying to build an exact match filter list like this.

Q. Thank you, Mr. Clark.

I see a list of eight rows of eight types of mitigations on this document. You have called this document on several occasions an exploratory document. So I would like to have Meta tell me, which of these eight mitigations were not implemented?

A. Okay. Let me read through them again real quick for that. We can go through them one by one.

Number one, deduplication of datasets was adopted as a mitigation. And there's a mitigation as a requirement from the dataset review set process and before data could

be used for training.

The statement "Single pass: We train with one epoch over IP rich data" -- and then in parentheses (common crawl) which has shown in academic literature to reduce memorization," that was adopted as a mitigation.

For clarity, because we were also measuring the likelihood of memorization doing just a single pass or single epoch either, that -- that expanded over time because we were able to demonstrate that it was lower numbers of epochs, not just a single pass. So it's slightly modified from how this is written.

The basic decoding methods where "We applied decoding methods (sampling, temperature) which statistically reduce likelihood of memorization," those are applied naturally in the -- or in the response side of the model. They are not identified as specific mitigations.

And why you would set the decoding methods had more to do with prompt quality response or -- or response quality as opposed to

IP mitigations so it's not specifically identified as an IP mitigation, but we do have -- we do have optimized decoding methods that are not at zero.

Sampling and temperature have to do with the randomization of responses, and so we -- we don't want them fully random, but we don't want -- we don't want them under-randomized.

Number 4 is purchased data. We have added our own purchased data, human annotations, license data into the training mix. This -- from a IP mitigation perspective, this was not adopted as an IP mitigation.

We do have our own annotation data that we have added to models and depend on that for a variety of portions of the training data.

License data, we don't have any textual licensed data or purchased licensed data.

Targeted URL removal where it says "We can delete business risk data sources by URL from our pretraining date (i.e., ticketmaster, orbitz or booking.com), this one, the way it's

written is fairly poorly written.

What this is talking about is from datasets that are based on URLs, which a good example would be Common Crawl where Common Crawl stores the data by URL, what data came from Ticketmaster or what data came from Orbitz or booking.com.

And so in the processing of that training data, it's possible to say there are certain domains which we would not want to process. And so by targeted URL removal or the block list that is referred to the risks and blockers -- we had talked about the block list, I believe prior in my testimony. What this blocked list referred to were like Ticketmaster, Orbitz and booking.com specifically are all sites who -- who had strong enforcement around scraping of their sites.

And this is, once again being in March of '23, prior to Common Crawl being very transparent about what they crawled and didn't and how people could block Common Crawl from

crawling their data.

And so we filtered out sites that had -- when we processed the Common Crawl data, we filtered out the data from sites whose URLs matched against that block list because of how the data was scraped.

Number 6 is the hotfixed IP filter list.

So just to clarify on 5, but 5 was adopted but not in the form in which is it's currently written or labeled or titled.

6, the hotfixed IP filter list was not adopted as a mitigation.

7, "Generation time duplicate detection: build an n-gram index of the training set to search for n-gram overlap at generation time." And "generations over a particular threshold could be rejected."

This was not adopted as a mitigation, both from a technical feasibility perspective but also because we put investments into sampling of the data instead and measuring

memorization potential of the model end to end.

And then finally, "Dynamic sampling: based on model likelihood, reject select generations with the hope that" those "are more likely to match training corpus."

Once again, rejected because it's -- it's not clear that it would have made any difference, and we focused on the mitigations both from a pretraining perspective in addition to measuring the likelihood of -- of memorization from an output or from a reidentified -- reidentified ability perspective.

Q. Thank you, Mr. Clark.

What copyright/IP mitigations were adopted that are not on this list?

A. We had deduplication. Sorry. I have got to walk through them just to remember them all.

Single pass, we had measurement of memorization. We had content from the block list, being URLs that matched the block list for datasets who are based around specific URLs.

And I just need a moment to recollect.

Oh, and data that self-identified as pirated or stolen was to be removed or excluded from training. A lot of these -- actually all of these mitigations, except for the last one, originated from doing this for privacy and preventing the model from regurgitating or reidentifying information about private individuals.

And so these overlapped with that, and then the measurement and memorization being the final one.

Q. And in terms of the memorization issue, you mentioned that, you know, it's an exploratory document. And Kathleen can correct me if I am mischaracterizing your testimony, but I think your testimony was that Meta's current position is that -- is not that you can consider any memorization to be a copyright risk.

Is there an amount of memorization that Meta does consider to be a copyright risk?

MS. HARNETT: Object to form.

A. Once again, "memorization" is a very generic term. Some memorization is intentional, and so some data is trained to have a higher likelihood of replying the information. Is Abraham Lincoln a president? Do you expect the model to be able to know that?

The threshold that was put in place is that Llama 2 with low levels of memorization, that for future models or future training, that -- that future models could not have a higher level of memorization than what Llama 2 was. And so Llama 2 became the -- the lowest threshold for potential memorization.

Q. And here in this document, there is actually a definition of memorization. Could you look to that with the three bullets on the first page ending in 459?

A. Yes. And so this is under the definition heading that states, "On that note, we define memorization as" -- and then a first bullet -- "An uncommon phrase (as determined by

perplexity)."

Second bullet, "At least 50 words in length." And the third bullet being "With a 90%" match to the training data.

Q. Does Meta agree that that's the working definition of memorization?

A. No. Once again, the purpose of this document was a theoretical exploration of potential mitigations and potential definitions by a product manager and engineers that then fed into what -- what memorization could be.

This, once again, is kind of, like, I'm defining any memorization to be a copyright risk and memorization is a very generic term. And so this -- this does not what was -- this is not what was Meta's position, as far as memorization.

Q. So what is Meta's definition of memorization if it's not this?

MS. HARNETT: I object to the -- object to the form and the scope and also, foundation.

A. Memorization is -- memorization, as defined, is not these three items where it's an uncommon phrase or at least 50 words in length or with a 90 percent match to training data.

As it was defined, memorization is, once again, a broad technological concept. And so memorization is defined as the model's ability to repeat what is in training data as an exact match and is measured across various lengths -- not 50 words in length, but it's measured across a variety of the different lengths of phrases and random sets in order to do that testing.

And memorization -- memorization is generically defined as that ability and then -- yeah, that's -- there isn't a more specific definition.

BY MR. STEIN:

Q. I just don't understand the problem with this definition that's in this document.

What is Meta saying is incorrect about this definition?

MS. HARNETT: Objection to the form.

A. Sure. An uncommon phrase isn't measurable. It's not determined -- it has nothing to do with -- there are many uncommon phrases in training data. And the model in generating out -- outputs, both for how it's instructed to do so from a certain amount of randomness, but also how it is trained to do so can generate an uncommon phrase.

And so just the way this is worded is -- doesn't have anything to do with the likelihood of memorization or being memorization.

At least 50 words in length, as I stated, when we look for -- we have defined memorization to be the likelihood of the model outputting content that matches what is in training data. We don't just look at a specific length of data. We measure memorization across shorter and longer sets of phrases, numbers of words.

And then with the 90 percent match to the training data, once again, because the other two bullets here are either inaccurate or

ineffective or only, like, points in time in estimation for this explorative study, the -- the 90 percent match to training data doesn't match the model's ability to output training data exactly.

So these -- these three criteria are not part of what the application is in Meta's position, which is why I state, like, memorization, as defined, is the model's ability to output data exactly as it was in a model.

BY MR. STEIN:

Q. So I -- if I ask Llama to tell me the first paragraph of Stephen King's book The Stand, it will tell me what that first paragraph is. Is that memorization?

MS. HARNETT: Objection to the form and the scope. And this is not about the document.

MR. STEIN: Well, we're trying to understand the definition of memorization, which is the document. And he's telling me that this definition isn't -- isn't Meta's

around broad, publicly availability data, we -- you know, it's -- nothing is 100 percent, but we put the -- in place the mitigations to prevent as much memorization as possible.

BY MR. STEIN:

Q. Does Meta consider a removal of copyright information to a part of copyright IP mitigations?

MS. HARNETT: Object to the form. Vague and not related to the document.

MR. STEIN: Is there -- just --

MS. HARNETT: Can you tie that to the document?

BY MR. STEIN:

Q. The document is about copyright IP mitigations. So I'm wondering if Meta would consider the removal of copyright information or -- or copyright management information to be a copyright IP mitigation.

MS. HARNETT: Object to the form. Vague and not related to the document.

He is not here to testify about

mitigations generally. He is here to testify about this document.

MR. STEIN: Are you instructing him to not answer this particular question?

MS. HARNETT: He can answer the question because we want to get the deposition over, but I am just -- I am making my objection for the record that you are not asking questions about the document or directly related to the document.

A. We had discussed this prior in a prior deposition, and I did not prepare or have that document as a reference for all of the types of information. Copyright is as a phrase or the copyright symbol or the back/N or the similar things are repeatable sets of ASCII characters that were not removed because they were copyright material. They were removed because of the amount of times that that and other kinds of things were removed from the content because of how repeatable they were, how they existed in spots that would interrupt tokens and break

tokens.

And without doing that cleanup from a data parsing perspective, that would equal poor performance in the model and not allow the model to hit the industry benchmark. So extra line spaces, for instance, where somebody hits return five times, those five were removed so the paragraph ending is still there, but the extra things are not.

So it removes noise and it removes repeatability and it removes other things. Otherwise, the model would just regurgitate the things that were repeated the most.

Q. And in terms of copyright/IP mitigations, when Meta measures memorization, does it have to call back on the training datasets themselves?

MS. HARNETT: Object to form and beyond the scope and vague.

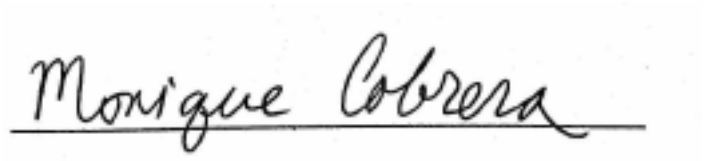
A. The -- as we walked through in prior testimony, the memorization process pulls random strings of ASCII characters in various lengths

CERTIFICATE OF SHORTHAND REPORTER

NOTARY PUBLIC

I, Monique Cabrera, the officer before whom the foregoing deposition was taken, do hereby certify that the foregoing transcript is a true and correct record of the testimony given; that said testimony was taken by me stenographically and thereafter reduced to typewriting under my direction; and that I am neither counsel for, related to, nor employed by any of the parties to this case and have no interest, financial or otherwise, in its outcome.

IN WITNESS WHEREOF, I have hereunto set my hand this 19th day of December, 2024.



MONIQUE CABRERA
Notary Public in and for the State of New York
County of Suffolk
My Commission No. 01CA6043156
Expires: 06/12/2026